Fine-Grained Textual Guidance for Generalized Multi-Modal Face Anti-Spoofing

Daiyuan Li, Zitong Yu, *Senior Member, IEEE*, Jinwu Hu, *Graduate Student Member, IEEE*, Guohao Chen, Jinghui Zeng, and Mingkui Tan, *Senior Member, IEEE*

***Abstract*—Multi-modal face anti-spoofing (FAS) is crucial for defending against presentation attacks in complex attack types and high-security scenarios. However, existing multi-modal FAS methods encounter two main limitations: 1) Most methods rely on classification supervision, which often fails to fully capture the distinctions between real faces and presentation attacks (PAs). 2) These methods depend solely on source domain data with limited PA types, leading to significant performance degradation when encountering unseen PA types and scenarios. To address these limitations, we propose a novel multi-modal fusion framework called Fine-grained Textual Guidance Multi-Modal Face Anti-Spoofing (FTG-FAS), which aligns natural language descriptions with multi-modal fused features to guide learning. Specifically, we propose a textual-guided token dropout module to select semantic invariant patch tokens for multi-modal fusion, thereby enhancing the model's generalization capability. In the testing phase, we propose FTG-FAS++, which leverages a self-distillation scheme with online source-free adaptation to further enhance model's performance in unseen scenarios. Specifically, we establish a teacher-student distillation framework, where the teacher model is fed with the complete image while the student model only receives masked tokens. During adaptation, we minimize the prediction discrepancy between the teacher and student in a unidirectional manner. Meanwhile, we propose a class-balanced sample selection strategy for stable source-free adaptation to prevent the model from overfitting to either real or spoof during the tuning process. Experiments show that FTG-FAS and FTG-FAS++ outperform SOTA methods by 6.91% and 8.72% in AUC on the cross-dataset leave-one-out protocols. Code will be available at https://github.com/iamcoming233/FTG-FAS.git**

***Index Terms*—Face anti-spoofing, multi-modal fusion, CLIP, source-free adaptation.**

Received 3 December 2024; revised 26 July 2025; accepted 7 September 2025. Date of publication 7 October 2025; date of current version 16 October 2025. This work was supported in part by the Joint Funds of the National Natural Science Foundation of China under Grant U24A20327, in part by the China Computer Federation (CCF)-Tencent Rhino-Bird Open Research Fund, and in part by Guangdong Provincial Key Laboratory under Grant 2023B1212060076. The associate editor coordinating the review of this article and approving it for publication was Dr. Pavel Korshunov. *(Daiyuan Li and Zitong Yu contributed equally to this work.) (Corresponding author: Mingkui Tan.)*

Daiyuan Li and Mingkui Tan are with the School of Software Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: ldyabroad@gmail.com; mingkuitan@scut.edu.cn).

Zitong Yu is with the School of Computing and Information Technology, Great Bay University, Dongguan 523000, China, also with Guangdong Provincial Key Laboratory of Intelligent Information Processing and Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen 518060, China, and also with Dongguan Key Laboratory for Intelligence and Information Technology, Dongguan 523000, China (e-mail: zitong.yu@ieee.org).

Jinwu Hu and Guohao Chen are with the School of Software Engineering, South China University of Technology, Guangzhou 510006, China, and also with Pazhou Laboratory, Guangzhou 510335, China (e-mail: fhujinwu@gmail.com, chenguohao987@gmail.com).

Jinghui Zeng is with the Shien-Ming Wu School of Intelligent Engineering and the School of Software Engineering, South China University of Technology, Guangzhou 511442, China, and also with Pazhou Laboratory, Guangzhou 510335, China (e-mail: zengjinghui0824@gmail.com).

Digital Object Identifier 10.1109/TIFS.2025.3613977

I. Introduction

FACE recognition [1], [2], [3] is one of the most widely implemented biometric technique across various authentication systems (*e.g.*, electronic payment, access control). However, face recognition systems are vulnerable to presentation attacks (PA), such as print, replay, and mask attacks, which pose significant security risks to identity verification systems [4]. To address this issue, Face Anti-Spoofing (FAS) [5], [6], [7], [8] have received extensive attention and research in both academia and industry. Uni-modal FAS methods [5], [9], [10], [11], [12], [13], [14] have achieved significant progress through the use of large, accumulated datasets [9], [15], [16], [17], [18]. However, the limited information provided by single-modal images makes it challenging to handle diverse PA types. In contrast, multi-modal FAS [16], [19], [20], [21], [22], [23], [24] methods leverage richer spoof cues, enabling models to more effectively address complex attack types and diverse application environments. The integration of multi-modal images has become a key strategy for improving the accuracy and robustness of FAS.

However, the discrimination and stability of multi-modal images vary across different scenarios, posing challenges for multi-modal face anti-spoofing. For instance, near-infrared (NIR) images can capture blood vessels and are relatively robust to variations in illumination, but lack rich color and texture details. Visible light images (RGB) provide abundant color and texture information but are highly sensitive to lighting changes [5], [10], [15]. Depth images can defend against planar attacks and coarse 3D attacks, but lack detailed surface texture and color information, making them less effective against more sophisticated attacks, such as realistic 3D face models or masks. Thus, how to extract and fuse the complementary information between different image modalities is crucial for multi-modal face anti-spoofing tasks.

Traditional multi-modal FAS methods [16], [20], [25], [26] concatenate or reweight the informative high-level features of different modalities (RGB, NIR, Depth) via attention modules (such as SEBlock [27]). However, these methods often

overlook the fact that multi-modal models tend to overfit to the dominant modality or spoof cues during the learning process [24], [28], resulting in suboptimal classification performance and poor generalization. To address this issue, George and Marcel [28] design a hard modality sample mining strategy that dynamically adjusts the loss weights of different modality branches to prevent the model from over-relying on any single modality. Liu et al. [8], [29] design cross-modal attention modules in the intermediate layers, allowing each modality to learn modality-independent knowledge from the remaining ones. However, training all parameters of the model is inefficient, and it leads to overfitting when only small-scale data is available. Inspired by partial parameter fine-tuning methods (*e.g.*, VPT [30], Adapter [31]) which focus on designing efficient tunable structures, recent studies propose parameter-efficient multi-modal frameworks for face anti-spoofing. Yu et al. [32] find that training all parameters of a large model on limited data may lead to catastrophic forgetting of general knowledge. Yu et al. [32], [33] introduce carefully designed parameter-efficient learnable structures to extract FAS-related local invariant features and reduce the forgetting of general knowledge. Lin et al. [24] propose an uncertainty estimation adapter module to suppress unreliable modality tokens and re-scale multi-modal gradients, which enhances stability and generalization.

Despite the significant progress made by existing multi-modal methods, they still suffer from critical limitations: 1) Most existing multi-modal FAS methods train the model using classification supervision, which fails to fully capture the intrinsic differences between real faces and spoof ones. 2) These methods only rely on source domain data to train a generalized FAS model. However, in real-world applications, the diversity of PA types and scenarios makes it difficult to address all unseen factors using limited source data alone. Although several domain adaptation methods have been developed in uni-modal studies [34], [35] that adapt models using source and unlabeled target data, their application in multi-modal FAS tasks is limited when source data is inaccessible due to potential storage and privacy constraints.

To address the above limitations, we propose a novel **F**ine-grained **T**extual **G**uidance Multi-Modal **F**ace **A**nti-**S**poofing method, called FTG-FAS. Specifically, unlike most multi-modal methods that use classification as supervision, we are motivated by vision-language models (VLMs) [36], [37], [38] to train a multi-modal model by aligning learnable textual descriptions with multi-modal image features in a shared feature space. Additionally, we design a Fine-grained Textual Guidance Cross-Attention Fusion Block to select more stable patch tokens for multi-modal fusion. This block drops unstable patch tokens based on the consistency of predictions between patch token embeddings and different fine-grained textual description features. To adapt our FTG-FAS to the unlabeled target domain under storage limitations and data privacy constraints, we propose a source-free multi-modal domain adaptation method via self-distillation and develop an active sample selection strategy. By analyzing the prediction distribution of the original model, we establish sample selection thresholds for reliable sample filtration, allowing our FTG-FAS to leverage only target domain data for unsupervised adaptation and thereby enhance the model's generalization capability in unseen domains.

We evaluate our FTG-FAS in leave-one-out protocol consisting of four multi-modal datasets: WMCA [39], PADISI [40], CASIA-SURF [16], and CASIA-CeFA [26], and achieve a **6.91**% improvement in AUC over the state-of-the-art methods. Furthermore, we adapt the FTG-FAS on 50% of the four unlabeled target domains using the proposed source-free domain adaptation method, referred to as FTG-FAS++. This enhanced version achieves an additional average improvement of **1.81**% over FTG-FAS.

In summary, the contributions of this paper are as follows:

- We propose a **F**ine-grained **T**extual **G**uidance Multi-Modal **F**ace **A**nti-**S**poofing (FTG-FAS). By leveraging the correlation consistency between image patch token embedding and textual description embedding of the same category but different attributes, we select semantic invariant patch tokens for fusion, thereby enhancing the generalization of the model.
- We design a source-free domain adaptation method to adapt our FTG-FAS without source domain data in a self-distillation manner. Moreover, we introduce a class-balanced sample selection strategy to identify reliable target domain samples for more stable online unsupervised adaptation. As a result, the enhanced version, FTG-FAS++, which integrates FTG-FAS with our source-free adaptation method, achieves improved performance in cross-domain scenarios.
- Experimental results on cross-domain benchmark including four cross-datasets demonstrate the effectiveness of our proposed FTG-FAS for multi-modal face anti-spoofing across scenarios and attack types. Additionally, we design multi-modal source-free adaptation experiments on target domain data. The FTG-FAS++ shows promising results in different unlabeled scenarios, further validating its effectiveness.

II. Related Works

A. Uni-Modal Face Anti-Spoofing

Traditional face anti-spoofing methods extract handcrafted local texture patterns (*e.g.*, SURF [41], LBP [15], HOG [42]) or capture micro-movement [43], [44], [45] from key parts of the face for classification. With the development of deep learning in vision tasks, Deep Neural Network (DNN) [46], [47], [48] is applied to solve the challenges in FAS tasks. Previous methods [49], [50] treat FAS as a binary classification task. However, binary classification-based methods may overfit arbitrary semantic information that is unrelated to FAS tasks. Therefore, more works focus on design complex auxiliary tasks (*e.g.*, pixel-wise supervision [5], [9], [10], [11], [12], [13], image generation [51], [52], [53], others [6], [54]) or specialized network structures [5], [10] to extract more intrinsic anti-spoofing features. Liu et al. [9] first propose to treat FAS as a pseudo face depth regression task. In subsequent research, various pixel-wise supervision (*e.g.*, binary mask [5], reflective surface [10], rPPG [55], [56]) are corporated into the FAS task.

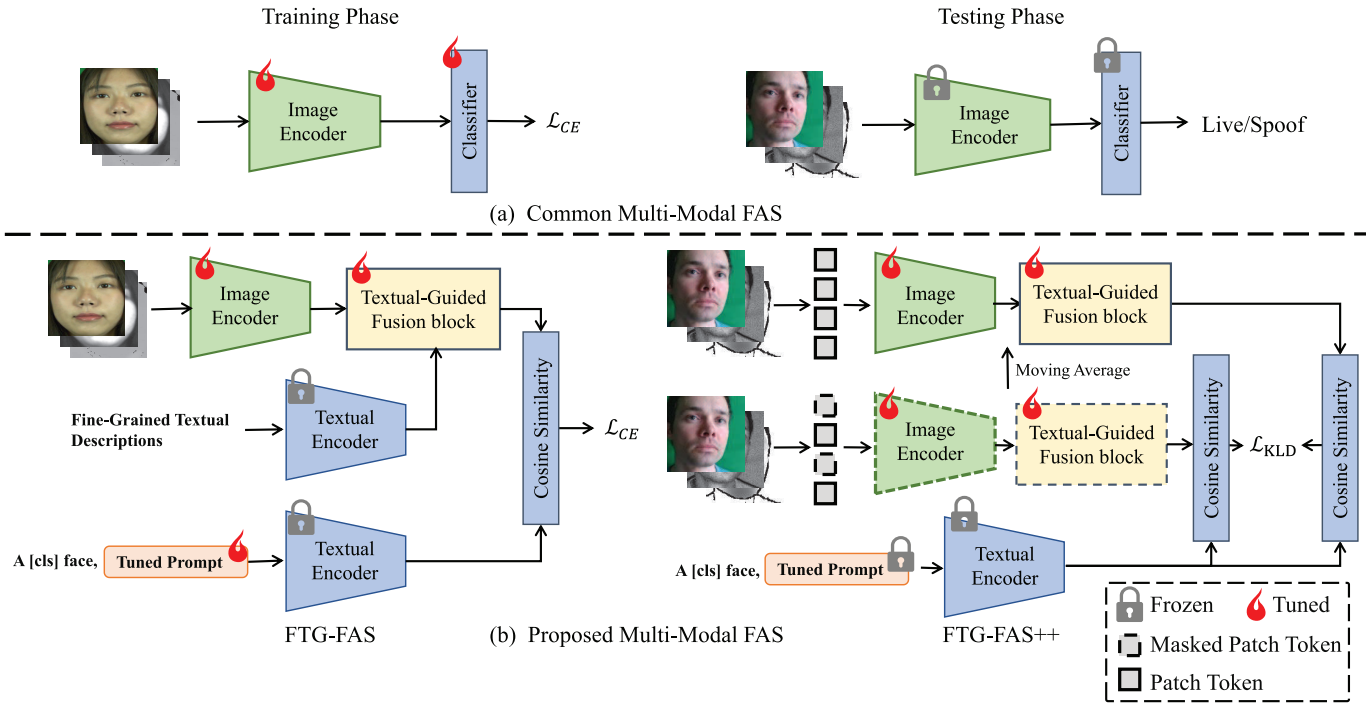


Fig. 1. Illustration of cross-domain scenarios in the context of multi-modal FAS. (a) Common pipeline for multi-modal face anti-spoofing. The model is trained with classification supervision, and its parameters remain fixed during the testing phase. (b) Proposed FTG-FAS and FTG-FAS++. In the training phase, FTG-FAS is trained by aligning the features of tuned textual prompts with the fused multi-modal image features. Additionally, fine-grained textual features are leveraged to guide the fusion of these multi-modal image features. In the testing phase, the pretrained FTG-FAS is adapted in self-distillation manner (referred to as FTG-FAS++).

Liu et al. [51] develop a fine-grained predictor by introducing a generation task that disentangles the spoofing trajectory map from the spoofing face. Wang et al. [6] design a patch-wise angular-margin softmax loss for fine-grained local spoofing cues mining. However, generation-based methods often lack stability, and supervised approaches generally require the introduction of additional models or operations. Yu et al. [5], [57] devise a Central Difference Convolution (CDC) [58] module to extract invariant local features and a CDC-based neural network structure search scheme. Wang et al. [12] design a learnable gradient operator in a data-driven manner for adaptive feature extraction.

Some recent works [14], [59], [60], [61], [62], [63], [64], [65] make an effort to address the domain-shift challenges that arise due to variations in environments, devices, *etc*. Jia et al. [59] perform single-side adversarial learning to learn a domain-invariant FAS feature space. Wang et al. [61] further extract general features by disentangling and shuffling the FAS-related and unrelated information in the face image. To tackle the catastrophic forgetting and unseen domain generalization problems, Cai et al. [66] propose a central difference convolutional adapter for a continual learning session. Srivatsan et al. [62] first propose aligning image representation with an ensemble of textual prompts to enhance the robustness of face anti-spoofing. Long et al. [67] propose Dual Sampling based Causal Intervention (DSCI) to eliminate the domain bias and improve generalization. Le and Woo [14] devise a sharpness-aware multi-domain optimizer for FAS task. Although uni-modal FAS methods have made significant progress, they have difficulty in handling complex application environments and diverse attack types. Therefore, multi-modal methods are gradually being widely adopted in practical applications.

B. Multi-Modal Fusion Face Anti-Spoofing

To address the drawback of uni-modal face anti-spoofing methods, multi-modal fusion studies focus on fusing complementary information from different sensors. Previously mainstream fusion face anti-spoofing methods [16], [19], [20], [68] are based on the attention mechanism to re-weight multi-modal features. For example, Zhang et al. [16] propose SD-Net that re-weight information channels of multi-modal images. Parkin and Grinchuk [68] and Kuang et al. [19] further introduce SD-Net-based multi-layer multi-modal fusion branches to enhance the contextual information cross-modalities. Wang et al. [20] propose a spatial and channel attention module to improve the differentiation of features between each modality. Shen et al. [69] employ patch-level input to extract spoofing-related differential features and introduce a modality feature erasure operation within the multi-modal features to prevent overfitting from different modalities. Liu et al. [26] propose a static-dynamic fusion mechanism applied to each modality and a partially shared fusion strategy to learn complementary features between modalities. George and Marcel [28] introduce a cross-modal focal loss to adjust the loss proportion of each modality, which helps learn complementary information between modalities while reducing the impact of overfitting.

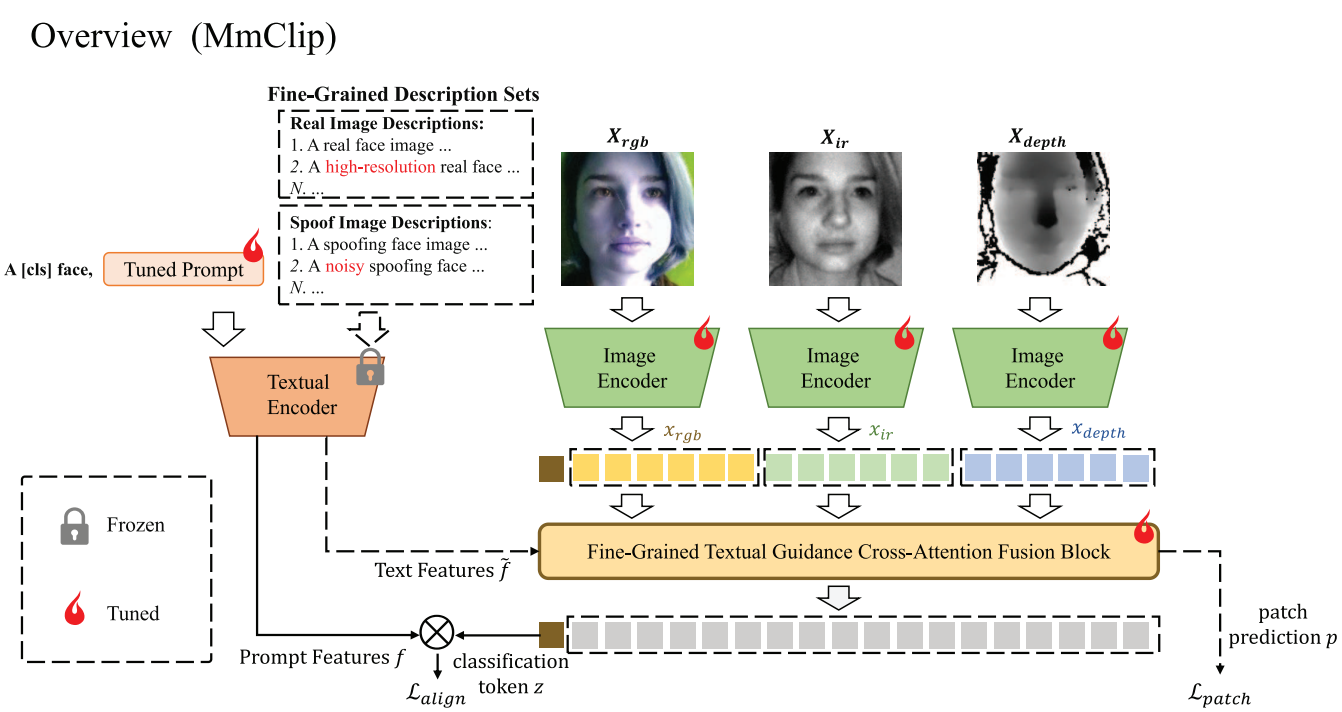


Fig. 2. The **F**ine-grained **T**extual **G**uidance Multi-Modal **F**ace **A**nti-**S**poofing (FTG-FAS) framework consists of a textual encoder, a set of Fine-grained textual description, an image encoder (shared across the three image modalities), and a *Fine-grained Textual Guidance Cross-Attention Fusion Block*. First, image patch features x are extracted using the image encoder, while tuned textual prompt features f and fine-grained description text features $\tilde{f}$ are obtained through the textual encoder. The features x and $\tilde{f}$ are then passed into the Fine-grained textual guidance Cross-Attention fusion block to obtain the classification token z and patch predictions p. The final prediction is calculated via Equ. (1). The image encoder and Fine-grained textual guidance Cross-Attention fusion block are trained by minimizing the $\mathcal{L}_{align}$, and the linear projection in Textual-Guided token dropout module is trained by minimizing the $\mathcal{L}_{\text{patch}}$.

Since the cross-attention mechanism can dynamically assign weights based on the input features, recent work has introduced cross-attention to enhance the multi-modal fusion performance. Wang et al. [70] combined a multi-modal structure of convolution and fully connected layers, taking into account both local and global feature fusion. Inspired by the success of self-attention in visual tasks, studies such as [8], [21], [22], and [23] have proposed cross-modal cross-attention feature fusion mechanisms. To mitigate the forgetting of general knowledge, [24], [32], [33], [71] proposed efficient parameter-efficient fine-tuning methods by inserting partially designed optimizable modules into the pretrained backbone. Lin et al. [24], [72] further designed a Monte Carlo uncertainty sampling adapter to erase unreliable modal tokens, enhancing the robustness of multi-modal fusion. Most recently, Liu et al. [23] proposed using tuned textual features to guide the learning of modal-independent image features. In contrast to their work, we aim to filter the relatively invariant image patch tokens by measuring the alignment consistency between patch embeddings and category-agnostic fine-grained textual descriptions.

C. Source-Free Domain Adaptation

Source-free domain adaptation [73], [74], [75], [76], [77] aims to address the domain shift problem when the source data and unlabeled target data are offset from the domain, and the source domain data are inaccessible. TTT [73], TTT++ [78] are designed to optimize the model with supervised branching and self-supervised objectives in the training phase, while only the self-supervised branching is used to update the model in the testing phase. These methods require introducing additional training tasks in the training phase, which is not flexible enough. Subsequent work directly utilizes pretrained models for test-time tuning. Wang et al. [74] first proposes to optimize the model prediction entropy at test time. Reference [79] uses predictions from multiple data augmentations of the same sample as a self-supervised signal implementation. Niu et al. [77] proposes a sharpness-sensitive test-time optimization method, which updates the model by filtering sharpness-insensitive samples with smaller gradient amplitudes to obtain more stable and robust results. Shu et al. [80] propose a test-time prompt tuning method on LVM models for the first time, adjusting only the Textual Prompt. Recently, some works devise source-free adaptation for a single modality. Liu et al. [81] proposes to separately aggregate the features of spoof and real faces from both the target and source domains and to align the predictive distributions between the target and source domains. TTSP [82] proposes to learn a set of bases that map samples from any unknown domain distribution to an existing set of pre-learned bases, enabling generalization during testing.

III. Proposed Method

In this work, we first propose a Fine-grained Textual Guidance Multi-Modal Face Anti-Spoofing framework (FTG-FAS) to fuse the stable multi-modal image features. As shown in Fig. 2, the FTG-FAS includes a multi-modal image encoder, a

textual encoder, a tuned text prompt, a set of Fine-grained textual Description, and a Fine-grained Textual Guidance Cross-Attention Fusion Block (*c.f.,* III-B). Furthermore, to reduce the performance degradation caused by domain distribution shifts in real-world applications, we devise a source-free domain adaptation method, which can adjust the model using only unlabeled multi-modal images of the target domain. Finally, we propose an enhanced version, called FTG-FAS++, by introducing the proposed source-free domain adaptation in FTG-FAS during the testing phase(*c.f.,* III-C).

A. Fine-Grained Textual Guidance Multi-Modal Face Anti-Spoofing

Following the generalized Contrastive Learning Image-Language Pretraining paradigm (CLIP) [36], the proposed FTG-FAS framework includes a multi-modal image encoder F_I, textual encoder F_T, and a fine-grained textual guidance cross-attention fusion block. As shown in Fig. 2. given multi-modal inputs $X = \{X_{rgb}, X_{ir}, X_{depth}\}$ including three modalities (RGB, NIR, Depth), we extract image features $x = \{x_{rgb}, x_{ir}, x_{depth}\}$ by F_I and take them as inputs of fine-grained textual guidance cross-attention fusion block B to obtain the classification token z. Moreover, due to the complex characteristics of face anti-spoofing samples, manually designing appropriate textual prompts for real or spoof faces is challenging. Inspired by CoOp [37], we employ a tunable textual prompt optimization method to tune textual prompts in the training phase. Specifically, we devise a pair of tuned prompts $T = \{T_l, T_s\}$ corresponding to real and spoof faces. The tuned prompt embeddings $f = \{f_l, f_s\}$ for T are calculated by F_T. The model parameters are trained to minimize the alignment loss between f and the corresponding image classification token z, as follows:

$$p = \frac{e^{(sim(z,f_l)/\tau)}}{e^{(sim(z,f_l)/\tau)} + e^{(sim(z,f_s)/\tau)}}, \tag{1}$$

$$\mathcal{L}_{align} = -logp \tag{2}$$

where $sim(\cdot,\cdot)$ denotes cosine similarity, f_l denotes the embedding of the real face textual prompt and f_s denotes the embedding of the spoof face textual prompt, τ is the temperature of the softmax. Notably, the image encoder parameters are trainable, while the textual encoder parameters remain fixed.

B. Fine-Grained Textual Guidance Cross-Attention Fusion Block

To effectively fuse patch features of multi-modal images, as shown in Fig. 3, we propose a Fine-grained Textual Guidance Cross-Attention Fuse Block. This block comprises a Textual-Guided Token Dropout module for selecting high-stability tokens and a multi-modal fusion module with H stacked self-attention and MLP layers.

1) Textual-Guided Token Dropout Module: Here, we aim to measure the stability of image patch tokens based on the consistency of similarity between the image patch token features and various fine-grained textual description features of the same class. Specifically, given a fine-grained textual description set containing N pairs of fine-grained descriptions for real face and spoof samples, we first extract real textual features $\tilde{f}_l$ and spoof textual features $\tilde{f}_s$ respectively using the textual feature encoder F_T. The x from three modalities are projected via $LinearProj(\cdot)$ to patch features $\tilde{x} \in \mathbb{R}^{L\times D}$. Then, we calculate the prediction $p \in R^{L\times N}$ between each image patch feature $\tilde{x}$ and, $\tilde{f}_l$, $\tilde{f}_s$ via Equ. (1). For the j-th image patch, we calculate the variance of its N corresponding predictions as its initial stability measure v_j, which is mathematically defined as:

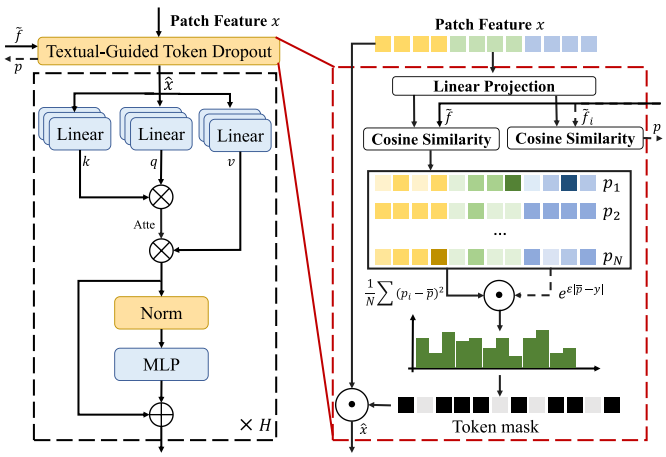


Fig. 3. **The detailed architecture of Fine-grained textual guidance Multi-Modal fusion block.** This block consists of a Textual-Guided token dropout module and a multi-modal fusion module composed of H cross-attention layers. The dashed arrows in the figure represent processes that occur only during training.

$$v_j = \frac{1}{\mathrm{N}} \sum_{i}^{N} (p_i^j - \bar{p}^j)^2 \tag{3}$$

where p_i^j is prediction of the j-th image patch feature corresponding to the i-th textual description, and $\bar{p}^j$ is the mean of all N predictions for the j-th image patch feature.

A larger v indicates that the patch token features are more inconsistent in predicting the same category, making the token relatively less stable. Moreover, selecting stable features solely based on prediction consistency may yield image tokens with higher consistency but lower informativeness (*i.e.*, less discriminative), as tokens might achieve low variance by consistently choosing the same answer. Therefore, in the training phase, we additionally introduce a weighting factor $e^{\varepsilon|\bar{p}^j - y|}$, where ε is a hyperparameter to control how token informativeness influences stability. The stability measure r_j for j-th image patch token can be further formulated as:

$$r_j = e^{\varepsilon|\bar{p}^j - y|} \cdot v_j \tag{4}$$

where y is the class label. According to Equ. (4), if the predicted score for a real face is closer to the label, its contribution is enhanced; otherwise, it is suppressed. Through the above steps, we retain the patch features that are relatively more informative. Finally, we set the patch features corresponding to the top γ largest r values to zero. Notably, we use only v as the stability measure in the testing phase.

2) Construction of Fine-grained Textual Description Set: We construct a set of fine-grained textual descriptions for both real and spoof faces images. Specifically, we introduce class-independent adjectives into the normal textual descrip-

Algorithm 1 The Training of FTG-FAS

Require: Multi-Modal Dataset $\mathcal{D} = \{X_{rgb}, X_{ir}, X_{depth}, y\}$, Tuned Prompt T, Textual Description $\tilde{T}$, Image Encoder F_I, Textual Encoder F_T, Multi-Modal Fusion Module M, Linear Projection $LinearProj$, Token Dropout Threshold γ.

while Not converged **do**
  for $(X_{rgb}, X_{ir}, X_{depth})$ in $\mathcal{D}$ **do**
    Calculate the patch features x_{rgb}, x_{ir}, x_{dep} for X_{rgb}, X_{ir}, X_{depth} via F_I.
    Calculate the textual features f, $\tilde{f}$ for T, $\tilde{T}$ via F_T.
    Obtain patch embeddings $\tilde{x}_{rgb}$, $\tilde{x}_{ir}$, $\tilde{x}_{dep}$ of x_{rgb}, x_{ir}, x_{dep} via $LinearProj(\cdot)$.
    Calculate outputs p_{rgb}, p_{ir}, p_{dep} between $\tilde{x}_{rgb}$, $\tilde{x}_{ir}$, $\tilde{x}_{dep}$ and one pair of text embedding $\tilde{f}_i$ with Equ. (1).
    Calculate alignment loss $\mathcal{L}_{patch}$ between p_{rgb}, p_{ir}, p_{dep} and y.
    Calculate outputs $\tilde{p}_{rgb}$, $\tilde{p}_{ir}$, $\tilde{p}_{dep}$ using Equ. (1) with $\tilde{x}_{rgb}$, $\tilde{x}_{ir}$, $\tilde{x}_{dep}$ and text embedding $\tilde{f}$.
    Calculate the stability measure r by Equ. (3) and Equ. (4).
    Obtain the final token $\hat{x}$ by setting the patch features with the top γ largest r values to zero.
    Input $\hat{x}$ into M to obtain classification token z.
    Calculate classification loss $\mathcal{L}_{align}$ between z and f.
    Update F_I, T, M and $LinearProj$ by minimize the Equ. (5).
  end for
end while

tion, primarily to describe image quality. (*e.g.* "noisy", "high-resolution", "blurred"). These adjectives, combined with normal descriptions, form fine-grained textual descriptions like "A **high-resolution** real face ... " and "A **blurred** real face ... ". In this work, we devise a fine-grained textual description set containing 7 pairs of real and spoofing textual descriptions.

3) Training of FTG-FAS: In the training phase, we jointly optimize the image encoder F_I, and Fine-grained Textual Guidance Cross-Attention Fusion Block M. Notably, when training the weights of $LinearProj(\cdot)$, we only select a pair of textual descriptions from the Fine-grained Textual Description Set. The features of this pair are then aligned with their corresponding patch token features, using the alignment loss denoted as $\mathcal{L}_{patch}$. The overall objective function is:

$$\mathcal{L}_{overall} = \mathcal{L}_{align} + \mathcal{L}_{patch} \tag{5}$$

The algorithm of training FTG-FAS is shown in Algorithm 1.

C. Class-Balanced Source-Free Adaptation for FTG-FAS

To mitigate the performance degradation caused by data domain shifts. Inspired by [74] and [77], we attempt to adapt the model using unlabeled target domain data in an unsupervised manner. Common approaches [74], [81] use a fixed entropy or prediction score threshold to filter reliable samples, followed by minimizing the prediction entropy of target samples for self-supervised training. However, this method has notable drawbacks: **a) Imbalance in Sample Selection**: Selecting samples based solely on entropy can result in an imbalanced number of real and spoof samples. **b) Extreme Prediction Values**: Optimizing for minimal entropy encourages the model to make overly confident predictions.

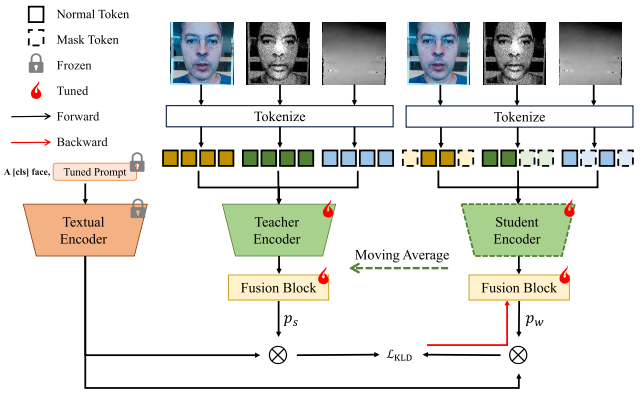


Fig. 4. Class-balanced Source-free adaptation. The teacher and student model are both copied from the pretrained model. We only update the parameters of layernorm in the student image encoder and fusion block by minimizing the KL loss between p_s and p_w. The teacher model is updated by exponential moving average using student model.

To address the aforementioned issues, as shown in Fig. 4, we design a self-distillation-based source-free method where the teacher model is used to guide the learning of the student encoder. Specifically, we first initialize both the teacher and student models with pretrained weights. For the teacher model, the complete patch tokens are input into the teacher encoder to generate the strong prediction p_s. we mask a portion of the input patch tokens by setting them to zero and feed the masked tokens into the model to generate the weak prediction p_w. In the testing phase, we minimize the Kullback-Leibler Divergence (KLD) between p_s (the strong branch prediction) and p_w (the weak branch prediction) to align the student's output with the teacher's output.

1) Class-Balanced Sample Selection for Source-Free Adaptation: To prevent the model from collapsing with all predictions in one class, we propose to select samples based on the predicted score distribution from the pretrained model. Specifically, given a target domain dataset of size S and a sample selection ratio θ, we construct a histogram using the pretrained model's probability distribution in the target domain. Then, we conduct a bidirectional traversal: starting from the highest-score bin, we accumulate sample counts until exceeding $\theta \times S$ (denoted as the high-region cumulative total Σ_{high}), and set this boundary value as ω_{high}. Simultaneously, we accumulate counts from the lowest-score bin until surpassing the Σ_{high}, and set this boundary value as ω_{low}. At this point, ω_{high} and ω_{low} serve as the high and low thresholds for sample selection.

In summary, the objective function of the proposed method can be formalized as:

$$\mathbb{I}(X) = \begin{cases} 1 & p_s > \omega_{\text{high}} \quad \text{or} \quad p_s < \omega_{\text{low}} \\ 0 & \text{else} \end{cases} \tag{6}$$

$$\mathcal{L}_{KLD} = \mathbb{I} \sum_{i=1}^{N} p_w \log \frac{p_w}{p_s} \tag{7}$$

TABLE I

COMPARISON WITH THE STATE-OF-THE-ART METHODS IN CROSS-DATASET LEAVE-ONE-OUT PROTOCOLS. WE REPORT THE AUC (%) AND HTER (%) OF THE PROPOSED FTG-FAS AND FTG-FAS++ IN FOUR EXPERIMENTS. THE **BOLD NUMBERS** INDICATE THE BEST RESULTS, WHILE THE UNDERLINED NUMBER IS THE SECOND-BEST RESULT

Method	C&P&S to W		C&S&W to P		P&C&W to S		P&S&W to C		Avg.	
	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑
MMDG [24]	12.79	93.83	18.95	88.64	15.32	92.86	29.93	76.52	19.25	87.96
VP-FAS [33]	16.26	91.22	21.76	85.46	24.42	81.07	39.35	66.55	25.45	81.08
MM-CDCN [25]	38.92	65.39	41.38	61.51	42.93	59.79	48.14	53.71	42.84	60.10
ViTAF [83]	20.58	85.82	30.75	73.03	29.16	77.80	39.75	63.44	30.06	75.02
AMA [32]	17.56	88.74	21.18	85.51	27.50	80.00	47.48	55.56	28.43	77.45
CMFL [28]	18.22	88.82	26.68	80.85	31.20	75.66	36.93	66.82	28.26	78.04
FTG-FAS (Ours)	3.58	99.50	**5.08**	**98.96**	14.03	93.22	20.81	87.78	10.88	94.87
FTG-FAS++ (Ours)	**3.46**	**99.58**	5.25	98.90	**9.39**	**96.77**	**16.33**	**91.46**	**8.61**	**96.68**

Algorithm 2 Class-Balanced Source-Free Adaptation for FTG-FAS

Require: Test Multi-Modal Samples $\mathcal{D}_{test} = \{X_{rgb}, X_{ir}, X_{depth}\}$, Tuned Textual Prompt T, Fine-grained Textual Descriptions $\tilde{T}$, Teacher Model F_t, Student Model F_s, Mask Threshold d.

Calculate the predictions for each samples in $\mathcal{D}_{test}$
Obtain the histogram of the predictions
Obtain the high threshold ω_{high} and low threshold ω_{low} from the histogram.
for $(X_{rgb}, X_{ir}, X_{depth})$ in $\mathcal{D}_{test}$ **do**
  Calculate the outputs p_s of X_{rgb}, X_{ir}, X_{depth} via F_t.
  Randomly mask d% patch tokens of X_{rgb}, X_{ir}, X_{depth}.
  Obtain the outputs p_w of masked tokens via F_s.
  Update student F_s by KLD loss in Equ. (7).
  Update teacher F_t by exponential moving average using F_s.
end for

where the function $\mathbb{I}(x)$ represents a selection function guided by data balance. The source-free adaptation algorithm is presented in Algorithm 2. In the testing phase, only the scale and bias parameters of the *LayerNorm* layers in the image encoder and multi-modal fusion block are updated. Moreover, the teacher model is updated via exponential moving average using the weights from the student model.

TABLE II

COMPARISON WITH THE STATE-OF-THE-ART METHODS ON 2 VS.2 PROTOCOLS. WE REPORT THE AUC (%) AND HTER (%) OF THE PROPOSED FTG-FAS AND FTG-FAS++ IN TWO EXPERIMENTS. THE **BOLD NUMBERS** INDICATE THE BEST RESULTS. THE UNDERLINED NUMBER IS THE SECOND-BEST RESULT

Method	C&W to P&S		P&S to C&W		Avg.	
	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑
MMDG [24]	20.12	**88.24**	36.30	70.35	28.21	79.29
VP-FAS [33]	25.90	81.79	44.37	60.83	35.14	71.31
ViTAF [83]	29.26	77.36	39.93	63.31	34.60	70.34
MM-CDCN [25]	29.28	76.88	47.00	51.94	38.14	64.61
AMA [32]	29.25	76.89	38.06	67.64	33.66	72.27
CMFL [28]	31.86	72.75	39.43	63.17	35.65	67.96
FTG-FAS (Ours)	21.93	84.41	34.07	70.72	28.00	77.57
FTG-FAS++ (Ours)	**18.85**	88.02	**33.11**	**72.22**	**25.98**	**80.12**

TABLE III

COMPARISONS OF THE MULTI-MODAL FAS METHODS IN THE MODALITY MISSING PROTOCOL

Methods	Missing D		Missing I		Missing D & I		Avg.	
	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑
MM-CDCN [25]	44.90	55.35	43.60	58.38	44.54	55.08	44.35	56.27
ViTAF [83]	34.99	73.22	35.88	69.40	35.89	69.61	35.59	70.64
AMA [32]	29.25	77.70	32.30	74.06	31.48	75.82	31.01	75.86
CMFL [28]	31.37	74.62	30.55	75.42	31.89	74.29	31.27	74.78
MMDG [24]	24.89	82.39	23.39	83.82	25.26	81.86	24.51	82.69
FTG-FAS(Ours)	**16.24**	**89.27**	**12.05**	**94.10**	**19.89**	**85.22**	**16.06**	**89.53**

IV. EXPERIMENTS

A. Datasets

We consider four different multi-modal face anti-spoofing datasets: WMCA (W) [39], PADISI (P) [40], CASIA-SURF (S) [16], and CASIA-CeFA (C) [26]. These datasets include RGB, IR, and Depth modalities for both real and spoofing samples and exhibit substantial domain gaps in presentation materials, capture devices, and illumination conditions.

- WMCA (W) [39] contains 2D (i.e. printed paper, replay) and 3D presentation attacks(i.e. paper mask, rigided mask, and silicon mask), with a total of 1679 video samples from 72 subjects. The RGB, near-infrared, and depth images are collected synchronously by Intel® RealSense™SR300, and the thermal channel is captured by Thermal CompactPRO. In this work, we use only RGB, near-infrared, depth data in WMCA.
- CASIA-SURF (S) [16] consists of 1000 subjects with 21000 videos, and contains paper, paper mask attacks.
- CASIA-CeFA (C) [26] contains 1607 subjects, collected from 3 different ethnicities (i.e., Africa, East Asia, and Central Asia) with four types of presentation attacks (i.e., photo, video replay, 3D mask, and silica gel face attacks) under different illumination conditions.
- PADISI (P) [40] contains 9 presentation attacks(*i.e.* printed paper, transparent mask, mannequin (fake head), silicone mask, half mask, makeup, tattoo, funny eye, and paper glasses from 360 subjects.

TABLE IV

ABLATION STUDY FOR ALL COMPONENTS OF FTG-FAS++ IN LEAVE-ONE-OUT PROTOCOLS. THE "TOKEN DROPOUT" MEANS THE MODEL THAT INTEGRATES OUR TEXTUAL-GUIDED PATCH TOKEN DROPOUT MODULE, AND "50% C-SFDA" MEANS THE MODEL TUNED USING 50% UNLABELED TARGET DOMAIN DATA. THE **BOLD NUMBERS** INDICATE THE BEST RESULTS

Method	C&P&S to W		C&S&W to P		P&C&W to S		P&S&W to C		Avg.	
	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑
Baseline	6.75	98.10	13.46	93.55	14.38	92.93	21.56	86.97	14.04	92.90
+ 50% C-SFDA	6.84	98.14	13.03	93.64	10.55	96.06	19.23	88.91	13.71	94.19
+ Token Dropout	3.58	99.50	**5.08**	**98.96**	14.03	93.22	20.81	87.78	10.88	94.87
+ token dropout + 50% C-SFDA	**3.46**	**99.58**	5.25	98.90	**9.39**	**96.77**	**16.33**	**91.46**	**8.61**	**96.68**

TABLE V

EFFECT OF THE NUMBER OF TARGET DATASETS ON LEAVE-ONE-OUT CROSS-DATASET PROTOCOL. WE REPORT THE AUC (%) AND HTER (%) OF FTG-FAS ++ IN SIX SAMPLING RATE SETTING. THE **BOLD NUMBERS** INDICATE THE BEST RESULTS

Rate for Adaptation	C&P&S to W		C&S&W to P		P&C&W to S		P&S&W to C		Avg.	
	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑
0% (w/o adaptation)	3.58	99.50	5.08	98.96	14.03	93.22	20.81	87.78	10.88	94.87
10%	3.46	99.59	**5.08**	**98.97**	11.96	94.97	18.76	89.47	9.82	95.75
30%	3.41	99.60	5.24	98.94	10.57	95.99	17.43	90.52	9.16	96.26
50%	3.46	99.58	5.25	98.90	**9.39**	**96.77**	**16.33**	**91.46**	**8.61**	**96.68**
70%	**3.15**	**99.66**	5.38	98.65	10.32	96.22	15.88	91.79	8.68	96.58
100%	4.60	99.36	5.52	98.85	10.06	90.38	18.52	90.81	9.68	96.23

B. Implementation Details

We implement the proposed method in PyTorch [84] and use ViT [85] as the backbone of the image encoder. In the training phase, we initialize the parameters of the image and textual encoder with the pretrained CLIP [36] models. All multi-modal images (RGB, IR, Depth) are resized to $224 \times 224 \times 3$ for input. For Depth images, the single channel is repeated across three channels. The learning rate is $5 \times \mathbf{10}^{-6}$ and the weight decay is 0.05. We train the networks for 70 epochs with a batch size of 48. We apply label smoothing with a factor of 0.1. The token dropout rate is set to 0.1. The learning rate is halved every 30 epochs. During source-free adaptation, 50% of the unlabeled target data are used for unsupervised training. The learning rate is $1 \times \mathbf{10}^{-4}$, and the batch size is set to 16. The input mask ratio is set to 0.4. The smooth factor of exponential moving average is set to 0.95. Other hyper-parameters involved are as follows: $\varepsilon = 2.7$, $\theta = 0.025$, $\tau = 1$, $H = 2$.

C. Performance Metrics

We use the Half Total Error Rate (HTER) and Area Under the Curve (AUC) metrics to evaluate the performance in cross-dataset experiments. HTER is the average of APCER and BPCER. Specifically, APCER (Attack Presentation Classification Error Rate) represents the proportion of spoof attack samples that are incorrectly classified out of the total number of spoof samples. BPCER (Bona Fide Presentation Classification Error Rate) represents the proportion of real face samples that are incorrectly classified out of the total number of real samples. AUC measures the area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate against the false positive rate at different threshold values.

D. Comparison Experiments

Following MMDG [24], we first evaluate our method in leave-one-out protocol, which three out of the four datasets are used for training, and the remaining dataset is used for testing in each experiment. For instance, C&P&S to W indicates that CASIA-CeFA, PADISI, and CASIA-SURF are the training sets, while WMCA is the test set. As shown in Tab. I, our proposed FTG-FAS achieves the highest average AUC and lowest HTER, outperforming state-of-the-art multi-modal FAS methods by 8.37% (19.25% to 10.88%) in average HTER and 6.91% (87.96% to 94.87%) in average AUC. Furthermore, we report the AUC and HTER of FTG-FAS++, which integrates the proposed source-free adaptation method with FTG-FAS. In all 4 experiments, the FTG-FAS++ outperforms the FTG-FAS by 1.81% (94.87% to 96.68%) in AUC and by 2.27% (10.88% to 8.61%) in HTER, demonstrating the effectiveness of our proposed source-free adaptation method.

We further compare FTG-FAS with the state-of-the-art methods in the 2 vs. 2 protocols, where two datasets are used for training and the remaining two are used for testing. As illustrated in Tab. II, the FTG-FAS suppressed by MMDG [24], of 1.81% (20.12% to 21.93%) in HTER and 3.83% (88.24% to 84.41%) in AUC on the C&W to P&S. However, the FTG-FAS outperforms the MMDG by 2.23% (36.30% to 34.07%) in HTER and 0.37% (70.35% to 70.72%) in AUC on P&S to C&W. Moreover, the enhanced version, FTG-FAS++, demonstrates consistent performance improvements across both two experiments. Specifically, FTG-FAS++ surpasses FTG-FAS with an average increase of 2.55% (77.57% to 80.12%) in AUC and a 2.02% (28.00% to 25.98%) reduction in HTER. Notably, FTG-FAS++ achieves the best performance in both experiments under the 2 vs. 2 protocol, highlighting its effectiveness.

Moreover, we evaluate FTG-FAS's cross-dataset performance under modality missing protocols. Following the

TABLE VI

COMPARISONS OF THE SOURCE-FREE ADAPTATION METHODS IN THE LEAVE-ONE-OUT PROTOCOL. WE REPORT THE AUC (%) AND HTER (%) FOR THE MODEL ADAPTED USING DIFFERENT SOURCE-FREE ADAPTATION METHODS. THE **BOLD NUMBERS** INDICATE THE BEST RESULTS

Method	C&P&S to W		C&S&W to P		P&C&W to S		P&S&W to C		Avg.	
	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑	HTER↓	AUC↑
FTG-FAS (Ours)	3.58	99.50	5.08	98.96	14.03	93.22	20.81	87.78	10.88	94.87
+ Tent [74]	8.01	96.68	5.08	98.96	49.29	51.55	26.95	78.46	22.33	81.41
+ Zero [86]	3.71	96.93	14.48	93.39	20.99	86.71	50.00	69.92	22.29	86.73
+ Cotta [79]	4.51	99.50	4.48	99.14	17.60	90.36	19.91	88.61	11.63	94.40
+ SAR [77]	3.59	99.30	5.08	98.98	17.40	90.33	19.92	86.61	11.50	93.81
+ EATA-C [87]	3.67	99.57	5.08	98.96	15.05	92.34	20.30	92.34	11.03	94.77
+ TPT [80]	3.59	99.56	5.08	98.96	14.06	93.23	20.78	88.76	10.88	95.13
+ SDA-FAS [88]	**3.36**	99.55	**3.28**	**99.28**	18.86	89.62	23.04	83.86	12.14	93.08
+ Tent & COME [89]	3.58	99.57	5.08	93.27	13.97	93.27	20.86	87.75	10.87	94.88
FTG-FAS++ (Ours)	3.46	**99.58**	5.25	98.90	**9.39**	**96.77**	**16.33**	**91.46**	**8.61**	**96.68**

TABLE VII

ABLATION STUDIES OF DIFFERENT TEXTUAL DESCRIPTIONS SETTING

Textual description setting	Avg.	
	HTER↓	AUC↑
Normal textual descriptions	13.53	93.47
Fine-grained textual descriptions (Ours)	**10.88**	**94.87**
five pairs	13.01	93.13
six pairs	11.34	94.69
seven pairs (Ours)	**10.88**	**94.87**

TABLE VIII

EFFECTIVENESS OF TEXTUAL DESCRIPTORS: PERFORMANCE COMPARISON BETWEEN FLIP [62] PROMPTS AND AUGMENTED VARIANTS WITH IMAGE QUALITY ADJECTIVES (DENOTED AS FLIP*)

Description	AUC(%)	HTER(%)
FLIP [62]	93.46	13.10
FLIP*	**93.50**	**12.67**

modality missing protocol in MMDG [24], we compare with multi-model methods across three modality missing scenarios. As shown in Tab. III, the proposed FTG-FAS consistently outperforms existing methods in all scenarios. Specifically, it achieves the lowest average HTER (16.06%) and highest AUC (89.53%) among the three scenarios. These results significantly surpass the state-of-the-art method MMDG by 8.45% in HTER and 6.84% in AUC, demonstrating FTG-FAS's superior robustness and effectiveness in handling incomplete multi-modal inputs.

E. Ablation Studies

1) Effectiveness of Framework Components: We evaluate all components of the proposed method, including the textual-guided patch token dropout module (abbreviated as Token Dropout in Tab.IV) and class-balanced source-free domain adaptation (abbreviated as C-SFDA in Tab.IV). The baseline model refers to the FTG-FAS with the textual-guided token dropout module removed. As shown in Tab. IV, incorporating the textual-guided patch token dropout consistently enhances the baseline model's performance across all four experiments under the leave-one-out protocol. The token dropout module achieves the largest performance improvement on the C&S&W to P setting, with a 5.41% (93.55% to 98.96%) improvement in AUC and a 8.38% (13.46% to 5.08%) reduction in HTER compared to the baseline model. On average result of four experiments, the Token Dropout yields a 1.97% (92.90% to 94.87%) increase in AUC and a 3.16% (14.04% to 10.88%) decrease in HTER. Furthermore, equipping the baseline model with both the token dropout module and C-SFDA results in further optimization of performance in the target domain. This combination of two components yields an average improvement of 1.81% (94.87% to 96.68%) in AUC and 2.27% (10.88% to 8.61%) in HTER. However, C-SFDA slightly degrades performance in the C&S&W to P experiments, This is likely because the pre-trained P-domain model already achieves class equilibrium, which diminishes FTG-FAS++'s core advantage in handling imbalance.

2) Effectiveness of Textual Descriptions: Here, we devise two protocols to investigate the impact of the construction method and the number of textual descriptions on the model performance. In the first protocol, we create a normal textual description set including 7 pairs of textual descriptions without descriptive adjectives (e.g., "this is a picture of a real face","‘ A photo of spoofing image"). We then compare the performance of models that apply the fine-grained and normal textual description set. In the second protocol, we compare 3 different settings varying number of fine-grained descriptions. Specific fine-grained textual descriptions are shown in Tab IX, where the first 5, 6, and 7 rows correspond to the respective settings. As shown in Tab. VII, the fine-grained textual descriptions outperform the normal ones, achieving a 2.65% (13.53% to 10.88%) reduction in HTER and a 1.40% (93.47% to 94.87%) increase in AUC. In the experiment on the number of textual descriptions, our FTG-FAS achieves optimal performance when using seven different pairs of textual descriptions. Furthermore, we conduct an extended evaluation of our fine-grained textual description strategy. Specifically, we first adopt the baseline textual prompts from FLIP [62], then construct enhanced variants by incorporating image quality descriptors (designated as FLIP*). The detailed descriptions of FLIP* are provided in Tab. X. As shown in Tab. VIII, FLIP* achieves a 0.04% absolute improvement in AUC (improving to 93.50%) and reduces HTER by 0.43% (decreasing to 12.67%) for face

TABLE IX

EXAMPLES OF DESCRIPTIONS. EACH ROW REPRESENTS A PAIR OF DESCRIPTIONS FOR A REAL FACE AND A SPOOFED FACE IMAGE. THE FIRST SEVEN PAIRS ARE THE PROPOSED FINE-GRAINED TEXTUAL DESCRIPTIONS, WHILE THE REMAINING PAIRS ARE NORMAL TEXTUAL DESCRIPTIONS

Textual descriptions for real face	Textual descriptions for spoofing face
a real face image with lifelike textures	a spoofing face image with replicated texture trace
a high-resolution real face image with lifelike texture	a low-resolution spoofing face image with replicated texture traces
a low-resolution real face image with lifelike texture	a high-resolution spoofing face image with replicated texture traces
a clear real face image with lifelike texture	a noisy spoofing face image with replicated texture traces
a noisy real face image with lifelike texture	a clear spoofing face image with replicated texture traces
a distorted real image with lifelike texture	a distorted spoofing face image with replicated texture traces
a complete de-spoofed face image with few replicated texture traces	a repaired spoofing face image with replicated texture traces
a photo of a real face	a pictaure of a fake face.
a picture of an authentic face	an image of a counterfeit face
a image of a true human face	a picture of a manipulated face
a snapshot of a real-life face	an image of a fabricated face
a photograph of a genuine face	a portrayal of a spoofed face
an actual photo of a face	a representation of a forged face
a real-life portrait of a face	a picture of a fake facial image

TABLE X

EXAMPLES OF DESCRIPTIONS FOR FLIP*

Textual descriptions for real face	Textual descriptions for spoof face
This is an example of a real face	This is an example of a spoof face
This is an example of a low-quality bonafide face	This is an example of a high-quality attack face
This is a real high-resolution face	This is not a low-resolution real face
This is how a high-resolution real face looks like	This is how a low-resolution spoof face looks like
A photo of a real face	A photo of a spoof face
This is not a clear spoof face	A printout shown to be a grainy spoof face

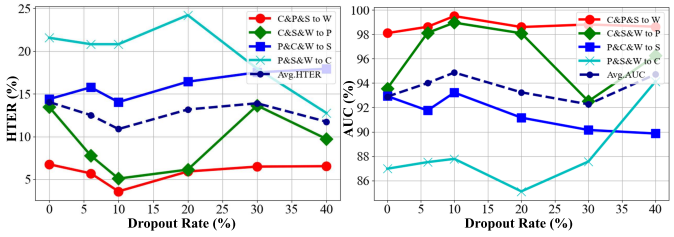


Fig. 5. Ablation study of the dropout rate $\gamma(\%)$ for textual-guided patch token dropout module. We report the AUC (%) and HTER (%) when $\gamma(\%)$ set to [6, 40].

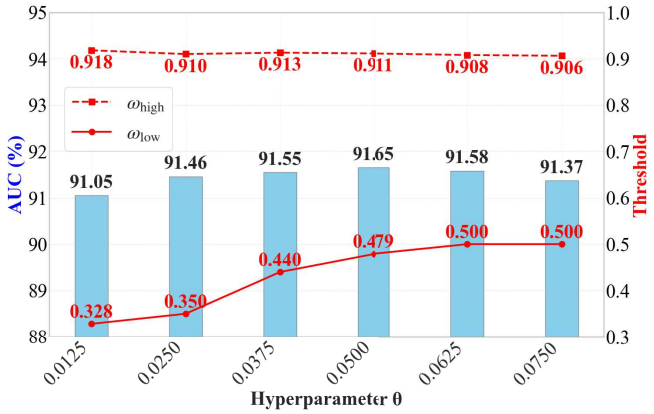


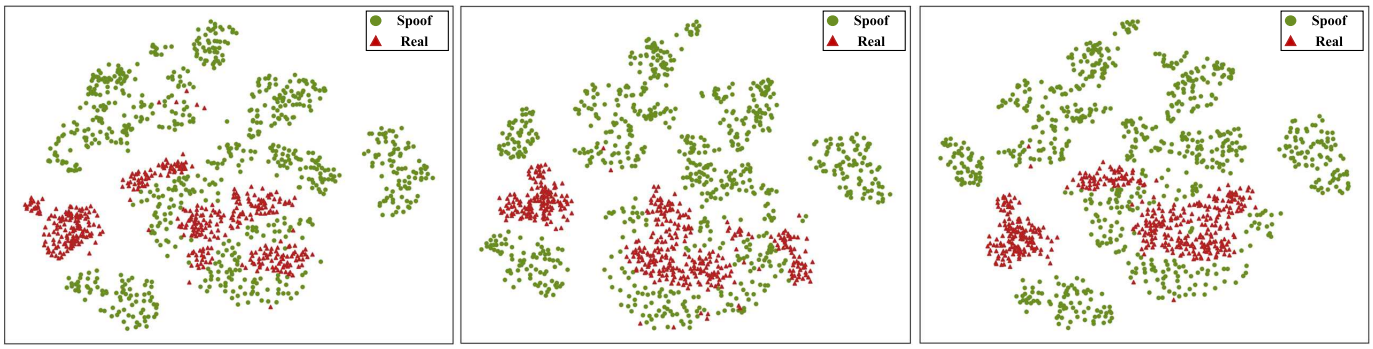
Fig. 6. Effect of θ to high threshold ω_{high}, low threshold ω_{low} and performance on P&S&W to C experiments.

anti-spoofing tasks. This improved performance demonstrates that explicitly incorporating image quality descriptors effectively enhances the model's discrimination capability against spoof attacks in unseen scenario.

3) Dropout Rate γ in Textual-guided Image Patch Dropout Module: As shown in Fig 5, we report the HTER (left figure) and AUC (right figure) in the different token dropout rates. The FTG-FAS achieves the best performance when the $\gamma(\%)$ is set to 0.1. Moreover, the FTG-FAS outperforms the baseline ($\gamma = 0$) in terms of average HTER and AUC when the threshold $\gamma = 0.06, 0.1, 0.2, 0.4$.

4) Effect of ε: We evaluate the impact of the weighting factor ε in Equ. (4) on the performance using leave-one-out protocols, and report the average HTER(%)↓. As shown in Tab. XII, the absence of weighting ($\varepsilon = 0$) leads to the highest HTER (14.63%), confirming that relying solely on prediction consistency discards informative tokens. The model achieves the best performance when $\varepsilon = 2.7$ (10.88% vs 14.63%). However, excessive weighting ($\varepsilon = 3.7$) diminishes the stability measure's effectiveness, degrading performance to 14.01% HTER.

5) Number of Samples for Source-free Adaptation: Here, we compare the model's performance after fine-tuning with 10%, 30%, 50%, 70%, and 100% of the target domain data, respectively. As shown in Tab. V, applying the proposed source-free adaptation method to tune the model on the target dataset positively impacted average AUC and HTER across most proportion settings. However, in the C&S&W to P exper-



(a) Baseline (b) Visualization of FTG-FAS (c) Visualization of FTG-FAS++

Fig. 7. The t-SNE [90] visualization of the fused multi-modal features. The left figure illustrates the fused feature of the model without the proposed fine-grained token dropout module (Baseline). The middle and right figures visualize the fused feature of the model without source-free adaptation and with proposed class-balanced source-free adaptation in protocol P&S&W to C respectively. The fusion feature clusters together when cooperating with the class-balanced source-free adaptation method.

iments, the improvement was less significant. Furthermore, the result in Tab V shows that continuously increasing the number of target domain data does not consistently improve model performance. For instance, in the P&C&W to S experiment, the model achieves the best performance in 50% number of the target data, but performance decreased when the proportion of target data was increased to 70%. Overall, after applying the proposed source-free domain adaptation in the target data, the average performance across four experiments shows an improvement over the baseline method.

6) Effectivness of Sample Ratio θ: We evaluate our proposed source-free domain adaptation method with different θ selected from $\{0.0125, 0.0250, 0.0375, 0.0500, 0.0625, 0.0750\}$ in P&S&W to C experiment, and illustrate the variations of the high threshold ω_{high}, low threshold ω_{low} and AUC in Fig. 6. The proposed method achieves the best performance with an AUC of 91.65% at $\theta = 0.05$, outperforming the default setting $\theta = 0.025$ by 0.19%. Minimal performance variation occurs for $\theta \in [0.025, 0.075]$, demonstrating robustness and low sensitivity to this hyperparameter within this range. Notably, while ω_{high} shows minor fluctuations, ω_{low} rises rapidly from 0.0325 to 0.05, indicating a positive prediction bias in the pretrained model.

7) Effectiveness of Source-free Domain Adaptation Methods: We compare proposed source-free domain adaptation method with other source-free adaptation methods in FTG-FAS, including general methods Tent [74], Cotta [79], SAR [77], TPT [80], Tent with COME [89], EATA-C [87], Zero [86] and uni-modal Source-Free FAS method, SDA-FAS [88]. All comparison results are based on the code provided in the respective paper. The proportion of the target domain dataset for source-free tuning is set to 50%. As shown in Tab. VI, the proposed method achieves superior AUC and HTER compared to other source-free adaptation methods in the C&P&S to W, P&C&W to S, and P&S&W to C experiments. The proposed method outperforms the best-performing method, TPT, by 1.55% (95.13% and 96.68%) in AUC, and 2.27% (10.88% and 8.61%) in HTER, respectively. Notably, some of the existing methods significant performance degradation compared to the pretrained model (FTG-FAS) in most experiments. For instance, Tent experiences substantial performance loss in P&C&W to S, with the model collapsing during adaptation, causing predictions to become heavily biased toward a single class. Similarly, in the P&C&W to S and C&P&S to W experiments, SAR, SDA-FAS, and Cotta show substantial negative optimization effects.

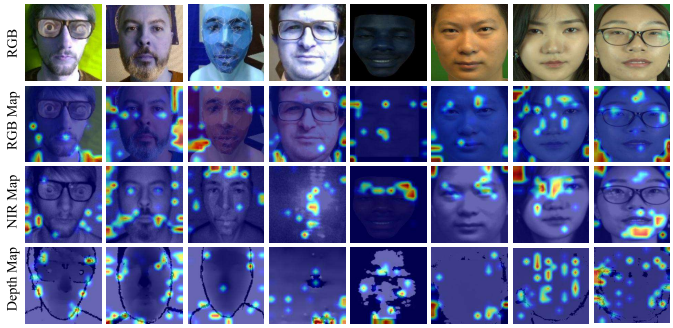


Fig. 8. Illustration of the dropped patch by textual-guided token dropout module. The highlighted areas denote patch tokens with a high probability of being discarded.

8) Mask Ratio d in Source-free Domain Adaptation: In this section, we examine the effect of ratio d in five different settings. As shown in Tab. XI. we report the average HTER of all four experiments in the leave-one-out protocol. The model's performance improves as the d increases within the interval [0, 0.4]. The best performance is achieved at $d = 0.4$, with an average HTER of 8.61%. However, performance does not improve further when the mask ratio exceeds 0.4, likely due to degradation caused by the loss of excessive masked information.

F. Visualization Analysis

1) Visual Analysis of Textual-guided Patch Token Dropout: To further analyze the effect of the textual token dropout module, we calculate the variance of the textual features with the token embedding of image patches from different modalities. We assign different brightness values to the image patch regions based on the variance of N predictions. A heatmap is then created by mapping these brightness values onto the original image, where redder colors indicate higher variance, signaling lower stability in those regions. As shown in Fig. 8, the first row displays the original RGB images, while the second row displays the corresponding heatmaps for the RGB images. The third and fourth rows are the heatmaps of the

TABLE XI
AVERAGE HTER(%) WITH DIFFERENT INPUT PATCH TOKEN MASK RATIOS IN LEAVE-ONE-OUT PROTOCOL

Mask ratio	0	0.1	0.2	0.3	0.4	0.5
HTER (%)	10.88	9.34	9.09	8.75	**8.61**	9.07

TABLE XII
AVERAGE HTER(%) WITH DIFFERENT ε

ε	0.0	0.7	1.7	2.7	3.7
HTER (%)↓	14.63	13.28	12.41	**10.88**	14.01

NIR and depth images, respectively. Most of the high response (low stability) regions in rgb heatmaps are located in the background of the image. However, some high-response areas in the NIR image are found in the face (forehead, eyes, and specular reflections), likely due to less background variation and more prominent changes in the hair and glasses regions. High response regions in depth maps are mostly located on the edges of facial contours.

2) Visualization of the Fused Multi-modal Image Feature: We extract fused multi-modal image features by the fusion block in P&S&W to C setting and visualize the fused features using *t*-SNE [90]. As shown in Fig. 7, red triangles and green points represent the fused features of real and spoof face samples, respectively. The Fig. 7 (a) shows the visualization of the model without the proposed textual-guided token dropout module (refer as "Baseline"). The Fig. 7 (b) and (c) display the visualization results of the FTG-FAS and FTG-FAS++ that apply the proposed class-balanced source-free adaptation. The clusters of real faces become noticeably more compact with the inclusion of the text-guided token dropout module in the fusion block, indicating improved class separability. Furthermore, as shown in Fig. 7 (c), the source-free adaptation method helps to group separated real samples more cohesively, demonstrating improved clustering of similar features.

V. CONCLUSION

In this work, we propose a fine-grained textual guidance multi-modal fusion framework for face anti-spoofing. First, to select stable multi-modal tokens for fusion, we develop a token selection strategy according to the consistency between predictions fine-grained textual features and patch token embeddings, based on which we design a fine-grained text-guided multi-modal fusion block. Secondly, we propose a self-distillation source-free adaptation method, further improving the performance of the model in the target domain. Moreover, we design a class-balanced sample selection method to prevent model overfitting either real or spoof class during the adaptation process. This method enhances the model performance in the unseen domain. Extensive experiments in the cross-domain benchmark of four multi-modal datasets demonstrate the effectiveness of the proposed method. Future work will investigate a broader spectrum of attack (*e.g.* adversarial attacks) and develop flexible test-time adaptation mechanisms to enhance real-world applicability.

REFERENCES

[1] U. Arora, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 6, pp. 399–458, 2003.

[2] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. 14th Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 87–102.

[3] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 5962–5979, Oct. 2022.

[4] J. Bigun, H. Fronthaler, and K. Kollreider, "Assuring liveness in biometric identity authentication by real-time face tracking," in *Proc. IEEE Int. Conf. Comput. Intell. Homeland Secur. Pers. Saf.*, Mar. 2004, pp. 104–111.

[5] Z. Yu et al., "Searching central difference convolutional networks for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5294–5304.

[6] C.-Y. Wang, Y.-D. Lu, S.-T. Yang, and S.-H. Lai, "PatchNet: A simple face anti-spoofing framework via fine-grained patch recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20281–20290.

[7] Z. Yu, R. Cai, Z. Li, W. Yang, J. Shi, and A. C. Kot, "Benchmarking joint face spoofing and forgery detection with visual and physiological cues," 2022, *arXiv:2208.05401*.

[8] A. Liu et al., "FM-ViT: Flexible modal vision transformers for face anti-spoofing," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 4775–4786, 2023.

[9] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 389–398.

[10] Z. Yu, X. Li, X. Niu, J. Shi, and G. Zhao, "Face anti-spoofing with human material perception," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 557–575.

[11] Y. Qin, Z. Yu, L. Yan, Z. Wang, C. Zhao, and Z. Lei, "Meta-teacher for face anti-spoofing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6311–6326, Oct. 2022.

[12] C. Wang, B. Yu, and J. Zhou, "A learnable gradient operator for face presentation attack detection," *Pattern Recognit.*, vol. 135, Mar. 2023, Art. no. 109146.

[13] T. Zheng et al., "MFAE: Masked frequency autoencoders for domain generalization face anti-spoofing," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 4058–4069, 2024.

[14] B. M. Le and S. S. Woo, "Gradient alignment for cross-domain face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 188–199.

[15] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "OULU-NPU: A mobile face presentation attack database with real-world variations," in *Proc. 12th IEEE Int. Conf. Autom. Face gesture Recognit.*, May 2017, pp. 612–618.

[16] S. Zhang et al., "CASIA-SURF: A large-scale multi-modal benchmark for face anti-spoofing," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 2, no. 2, pp. 182–193, Apr. 2020.

[17] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel, "The replay-mobile face presentation-attack database," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2016, pp. 1–7.

[18] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2012, pp. 1–7.

[19] H. Kuang et al., "Multi-modal multi-layer fusion network with average binary center loss for face anti-spoofing," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 48–56.

[20] G. Wang, C. Lan, H. Han, S. Shan, and X. Chen, "Multi-modal face presentation attack detection via spatial and channel attentions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1584–1590.

[21] P. Deng, C. Ge, X. Qiao, H. Wei, and Y. Sun, "Attention-aware dual-stream network for multimodal face anti-spoofing," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 4258–4271, 2023.

[22] Z. Yu, A. Liu, C. Zhao, K. H. M. Cheng, X. Cheng, and G. Zhao, "Flexible-modal face anti-spoofing: A benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 6346–6351.

[23] A. Liu et al., "FM-CLIP: Flexible modal CLIP for face anti-spoofing," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 8228–8237.

[24] X. Lin et al., "Suppress and rebalance: Towards generalized multi-modal face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 211–221.

[25] Z. Yu et al., "Multi-modal face anti-spoofing based on central difference networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 650–651.

[26] A. Liu, Z. Tan, J. Wan, S. Escalera, G. Guo, and S. Z. Li, "CASIA-SURF CeFA: A benchmark for multi-modal cross-ethnicity face anti-spoofing," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jul. 2021, pp. 1178–1186.

[27] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[28] A. George and S. Marcel, "Cross modal focal loss for RGBD face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 7882–7891.

[29] A. Liu and Y. Liang, "MA-ViT: Modality-agnostic vision transformers for face anti-spoofing," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, L. D. Raedt, Ed., Jul. 2022, pp. 1180–1186, doi: 10.24963/ijcai.2022/165.

[30] M. Jia et al., "Visual prompt tuning," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 709–727.

[31] Y.-L. Sung, J. Cho, and M. Bansal, "VL-adapter: Parameter-efficient transfer learning for vision-and-language tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, Jun. 2021, pp. 5227–5237.

[32] Z. Yu, R. Cai, Y. Cui, X. Liu, Y. Hu, and A. C. Kot, "Rethinking vision transformer and masked autoencoder in multimodal face anti-spoofing," *Int. J. Comput. Vis.*, vol. 132, no. 11, pp. 1–22, Nov. 2024.

[33] Z. Yu, R. Cai, Y. Cui, A. Liu, and C. Chen, "Visual prompt flexible-modal face anti-spoofing," 2023, *arXiv:2307.13958*.

[34] R. Quan, Y. Wu, X. Yu, and Y. Yang, "Progressive transfer learning for face anti-spoofing," *IEEE Trans. Image Process.*, vol. 30, pp. 3946–3955, 2021.

[35] F. Jiang, Y. Liu, H. Si, J. Meng, and Q. Li, "Cross-scenario unknown-aware face anti-spoofing with evidential semantic consistency learning," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 3093–3108, 2024.

[36] A. Radford et al., "Learning transferable visual models from natural language supervision," 2021, *arXiv:2103.00020*.

[37] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, Jun. 2022, pp. 16816–16825.

[38] Q. Ye et al., "CAT: Investigating and enhancing audio-visual understanding in large language models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 10, pp. 8674–8690, Oct. 2025.

[39] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel, "Biometric face presentation attack detection with multi-channel convolutional neural network," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 42–55, 2020.

[40] M. Rostami, L. Spinoulas, M. E. Hussein, J. Mathai, and W. AbdAlmageed, "Detection and continual learning of novel face presentation attacks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 14831–14840.

[41] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face antispoofing using speeded-up robust features and Fisher vector encoding," *IEEE Signal Process. Lett.*, vol. 24, no. 2, pp. 141–145, Feb. 2017.

[42] B. Peixoto, C. Michelassi, and A. Rocha, "Face liveness detection under bad illumination conditions," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 3557–3560.

[43] J. Li, "Eye blink detection based on multiple Gabor response waves," in *Proc. Int. Conf. Mach. Learn. Cybern.*, 2008, pp. 2852–2856.

[44] L. Wang, X. Ding, and C. Fang, "Face live detection method based on physiological motion analysis," *Tsinghua Sci. Technol.*, vol. 14, no. 6, pp. 685–690, Dec. 2009.

[45] A. K. Singh, P. Joshi, and G. C. Nandi, "Face recognition with liveness detection using eye and mouth movement," in *Proc. Int. Conf. Signal Propag. Comput. Technol. (ICSPCT)*, Jul. 2014, pp. 592–597.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jun. 2016, pp. 770–778.

[47] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[48] D. Guo, K. Li, B. Hu, Y. Zhang, and M. Wang, "Benchmarking micro-action recognition: Dataset, methods, and applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 6238–6252, Jul. 2024.

[49] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," 2014, *arXiv:1408.5601*.

[50] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid, "An original face anti-spoofing approach using partial convolutional neural network," in *Proc. 6th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Dec. 2016, pp. 1–6.

[51] Y. Liu, J. Stehouwer, and X. Liu, "On disentangling spoof trace for generic face anti-spoofing," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 406–422.

[52] S. Liu et al., "Dual reweighting domain generalization for face presentation attack detection," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Montreal, QC, Canada, Aug. 2021, pp. 867–873.

[53] B. Zhang, X. Zhu, X. Zhang, and Z. Lei, "Modeling spoof noise by de-spoofing diffusion and its application in face anti-spoofing," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2023, pp. 1–10.

[54] K.-Y. Zhang et al., "Structure destruction and content combination for face anti-spoofing," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Mar. 2021, pp. 1–6.

[55] Z. Yu, X. Li, P. Wang, and G. Zhao, "TransRPPG: Remote photoplethysmography transformer for 3D mask face presentation attack detection," *IEEE Signal Process. Lett.*, vol. 28, pp. 1290–1294, 2021.

[56] C. Yao, J. Ren, R. Bai, H. Du, J. Liu, and X. Jiang, "Mask attack detection using vascular-weighted motion-robust rPPG signals," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 4313–4328, 2023.

[57] Z. Yu, J. Wan, Y. Qin, X. Li, S. Z. Li, and G. Zhao, "NAS-FAS: Static-dynamic central difference network search for face anti-spoofing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 9, pp. 3005–3023, Sep. 2021.

[58] Z. Su et al., "Rapid salient object detection with difference convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 10, pp. 1–16, Oct. 2025.

[59] Y. Jia, J. Zhang, S. Shan, and X. Chen, "Single-side domain generalization for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 8484–8493.

[60] G. Wang, H. Han, S. Shan, and X. Chen, "Unsupervised adversarial domain adaptation for cross-domain face presentation attack detection," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 56–69, 2021.

[61] Z. Wang et al., "Domain generalization via shuffled style assembly for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 4113–4123.

[62] K. Srivatsan, M. Naseer, and K. Nandakumar, "FLIP: Cross-domain face anti-spoofing with language guidance," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 19628–19639.

[63] X. Wang et al., "TF-FAS: Twofoldelement fine-grained semantic guidance for generalizable face anti-spoofing," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2024, pp. 1–14.

[64] X. Ge et al., "DiffFAS: Face anti-spoofing via generative diffusion models," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 144–161.

[65] Z. Wang et al., "Consistency regularization for deep face anti-spoofing," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1127–1140, 2023.

[66] R. Cai et al., "Rehearsal-free domain continual face anti-spoofing: Generalize more and forget less," 2023, *arXiv:2303.09914*.

[67] X. Long, J. Zhang, S. Wu, X. Jin, and S. Shan, "Dual sampling based causal intervention for face anti-spoofing with identity debiasing," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 851–862, 2024.

[68] A. Parkin and O. Grinchuk, "Recognizing multi-modal face spoofing with face recognition networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1617–1623.

[69] T. Shen, Y. Huang, and Z. Tong, "FaceBagNet: Bag-of-local-features model for multi-modal face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1611–1616.

[70] W. Wang, F. Wen, H. Zheng, R. Ying, and P. Liu, "Conv-MLP: A convolution and MLP mixed model for multimodal face anti-spoofing," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2284–2297, 2022.

[71] J. Yang et al., "DADM: Dual alignment of domain and modality for face anti-spoofing," 2025, *arXiv:2503.00429*.

[72] X. Lin et al., "Reliable and balanced transfer learning for generalized multimodal face anti-spoofing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 9, pp. 7608–7625, Sep. 2025.

[73] Y. Sun, X. Wang, Z. Liu, J. J. Miller, A. A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 9229–9248.

[74] D. Wang, E. Shelhamer, S. Liu, B. A. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *Proc. 9th Int. Conf. Learn. Represent.*, 2020, pp. 1–15.

[75] M. Boudiaf, R. Mueller, I. B. Ayed, and L. Bertinetto, "Parameter-free online test-time adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8334–8343.

[76] Y. Iwasawa and Y. Matsuo, "Test-time classifier adjustment module for model-agnostic domain generalization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 2427–2440.

[77] S. Niu et al., "Towards stable test-time adaptation in dynamic wild world," in *Proc. 11th Int. Conf. Learn. Represent.*, 2023, pp. 1–27.
[78] Y. Liu, P. Kothari, B. G. V. Delft, B. Bellot-Gurlet, T. Mordan, and A. Alahi, "TTT+: When does self-supervised test-time training fail or thrive?," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 21808–21820.
[79] Q. Wang, O. Fink, L. Van Gool, and D. Dai, "Continual test-time domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7201–7211.
[80] M. Shu et al., "Test-time prompt tuning for zero-shot generalization in vision-language models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 14274–14289.
[81] Y. Liu, Y. Chen, W. Dai, M. Gou, C. Huang, and H. Xiong, "Source-free domain adaptation with contrastive domain alignment and self-supervised exploration for face anti-spoofing," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 511–528.
[82] Q. Zhou, K.-Y. Zhang, T. Yao, X. Lu, S. Ding, and L. Ma, "Test-time domain generalization for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 175–187.
[83] H.-P. Huang et al., "Adaptive transformers for robust few-shot cross-domain face anti-spoofing," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 37–54.
[84] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–18.
[85] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[86] M. Farina, G. Franchi, G. Iacca, M. Mancini, and E. Ricci, "Frustratingly easy test-time adaptation of vision-language models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 129062–129093.
[87] M. Tan et al., "Uncertainty-calibrated test-time model adaptation without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 8, pp. 6274–6289, Aug. 2025.
[88] Y. Liu, Y. Chen, W. Dai, M. Gou, C.-T. Huang, and H. Xiong, "Source-free domain adaptation with domain generalized pretraining for face anti-spoofing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5430–5448, Aug. 2024.
[89] Q. Zhang, Y. Bian, X. Kong, P. Zhao, and C. Zhang, "COME: Test-time adaption by conservatively minimizing entropy," in *Proc. Int. Conf. Learn. Represent.*, 2024, pp. 1–25.
[90] L. V. D. Maaten and G. E. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

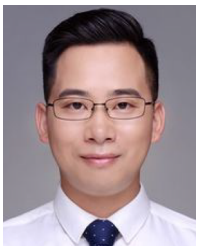
Daiyuan Li received the M.S. degree in computer software and theory from Hunan Normal University, China, in 2017. He is currently pursuing the Eng.D. degree with South China University of Technology, Guangzhou. His research interests include computer vision and face anti-spoofing.

Zitong Yu (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering from the University of Oulu, Finland, in 2022. He was a Post-Doctoral Researcher at the ROSE Laboratory, Nanyang Technological University. He was a Visiting Scholar at TVG, University of Oxford, from July 2021 to November 2021. He is currently an Associate Professor with Great Bay University, China. His research interests include human-centric computer vision and biometric security. He was a recipient of the IAPR Best Student Paper Award, the IEEE Finland Section Best Student Conference Paper Award in 2020, the Second Prize of the IEEE Finland Joint Chapter SP/CAS Best Paper Award in 2022, and the World's Top 2% Scientists by Stanford from 2023 to 2024.

Jinwu Hu (Graduate Student Member, IEEE) received the B.E. degree from Foshan University, Foshan, China, in 2020, and the M.S. degree from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2023. He is currently pursuing the Ph.D. degree with South China University of Technology, Guangzhou, China. He has published several journals/conference papers, including IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON MEDICAL IMAGING, IEEE TRANSACTIONS ON BIG DATA, ICML, IJCAI, CVPR, and ACM MM. His research interests include computer vision, machine learning, large language models, and reinforcement learning. He has served as a reviewer for many academic journals, including IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, NN, and PR.

Guohao Chen received the master's degree from the School of Software Engineering, South China University of Technology, Guangzhou, China, in 2025. He has published papers in top venues, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, NeurIPS, ICML, ACM MM, and IJCAI. His research interests include machine learning and mainly focus on inference-time learning. He has been invited as a reviewer for top-tier conferences and journals, including NeurIPS, ICLR, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.

Jinghui Zeng received the B.S. degree in optoelectronic information science and engineering from South China University of Technology, Guangzhou, China, in 2017, and the Ph.D. degree in software engineering from the Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, in 2025. His research interests include computer vision, 3-D measurement, and surface reconstruction.



Mingkui Tan (Senior Member, IEEE) received the bachelor's degree in environmental science and engineering and the master's degree in control science and engineering from Hunan University, Changsha, China, in 2006 and 2009, respectively, and the Ph.D. degree in computer science from Nanyang Technological University, Singapore, in 2014. From 2014 to 2016, he was a Senior Research Associate of computer vision at the School of Computer Science, The University of Adelaide, Australia. He is currently a Professor with the School of Software Engineering, South China University of Technology. His research interests include machine learning, sparse analysis, deep learning, and large-scale optimization.